# Sender-Receiver Games*

Ronald Peeters† Jos Potters‡

May 1999

**Abstract**

Standard game-theoretic solution concepts do not guarantee meaningful communication in cheap-talk games. In this paper, we define a solution concept which guarantees communication for a large class of games by designing a behavior protocol which the receiver uses to judge messages sent by the sender on acceptability. For that, we will make use of the Nash equilibrium concept for which truth-telling is a consequence. Uniqueness is nevertheless not a consequence, but after reasonable selection it is. Further, we will come to a method to compute all equilibria very easily.

*JEL classification: C72, D82.*

*Keywords:* Noncooperative game theory, Signalling, Sender-Receiver games.

## Introduction

In signalling games the importance of the fact that the parties understand what the signals mean is often underestimated. Many authors give a hint in the story companying their model but in the model itself this fact is very often suppressed. In Cho and Kreps' beer-and-quiche game for instance (see Cho and Kreps (1987)) the fact that 'drinking beer' is a signal of a 'strong type' can only be derived indirectly from the fact that the payoff for a strong type 'drinking beer' is higher than for a strong type 'eating quiche'. The story becomes less intriguing but clarity would have been served if the signals would have been 'I am strong/weak' under the condition that 'lying' is costly. If parties in a conflict try to

*We would like to thank P.Jean-Jacques Herings, Hans Reijnierse and Dries Vermeulen for their useful suggestions.

†Department of Econometrics and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, E-mail: R.J.A.P.Peeters@kub.nl

‡Department of Mathematics, Nijmegen University, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands, E-mail: potters@sci.kun.nl

give a signal by 'burning money' (Van Damme (1989), Hurkens (1995)) it should be clear what this signal means ('I do not frighten away from blackmailing you?') and that it has something to do with the decisions that follow. This is especially urgent, if the signal is not given but could have been given. In Crawford and Sobel (1982) it is mentioned in passing that the signal $y$ is a 'noisy estimate' of the real value of a variable $m$. By the way, the word 'noise' is here used in the unusual sense that the informed player causes the noise deliberately: he is telling the truth but not the whole truth. In *sender-receiver games* the messages sent by the sender are meant to convey information about the type of the sender to the benefit of both players. In the model the messages are just points in a message space and the only distinction between messages is the cost of sending the message. Very often this is not enough to convey any information.

*If every message can tell everything, no message is telling anything.*

Some recent papers on sender-receiver games (e.g. Rabin (1990), Farrell (1993), Zapater (1993)) recognize this fact and an a priori relation between a message and the information that can be transmitted by the message, is assumed: messages are like letters, they contain information. We see this as a necessary extension of the theory, not because of the little predictive power of the theory ('everything goes') but because it predicts something that will not (cannot) occur (without assuming that the receiver is 'clairvoyant' and then there is no need for messages). Another tendency one can observe in the more recent literature on signalling games is the choice for rationalizability concepts instead of Nash equilibrium concepts to explain behavior (see e.g. Hurkens (1995)). Also Rabin advocates this idea eloquently. In the theory followed in this paper we return, at a higher level, to Nash equilibrium behavior.

In this paper we will formulate a theory for communication between rational agents and we will use the theory in the paper of Rabin (1990) as a 'pièce de résistence' for our theory. Partially we will go along with Rabin's theory. We shall agree with him in the following points:

(1) Messages are bearers of information and not just points in a message space. There must be an *a priori relation between messages and types*. Messages are telling something about the type of the player. Rabin emphasizes that the same information can be conveyed by different messages and that mostly more natural language is used than prosaic statements as "My type is a type from $S$". In our opinion different messages conveying the same information can be identified without any problem. Not the phrasing of the information is important but the information itself. Moreover,

using more natural language often implies more ambiguity. Many interpretations of a message are possible. A message like "You should invest in my company" (Rabin) can interpreted as "Given my type, it is a good (better or the best?) action for you (or for me?) that you invest in my company". In this paper we will assume that only the prosaic but unambiguous messages "My type is a type from $S$" (in the sequel denoted by $[S]$) are available.

(2) The main issue is the formulation of behavior rules according to which the credibility of messages can be determined. In Rabin's theory this part is played by —what he calls— *credible message profiles*. We will introduce *acceptable message profiles* (AMP's) and *completely acceptable message profiles* (CAMP's). The idea behind both concepts is the same: it must be safe for the receiver of the message 'to believe the messages in the profile'.

(3) In Rabin's theory there is essentially only *one* credible message profile. In our theory we end up with a unique maximal CAMP for a generic class of games.

It is here that important differences between the two theories appear, namely:

(1') The definition of (completely) acceptable message profiles rests on Nash equilibrium behavior. It is *not* necessary that every (selected) message 'triggers' the best action for the types mentioned in the message but that it induces the best action for the types *sending the message* among the actions that can be induced by a message in the profile (under the assumption that the receiver believes the messages in the message profile and only these messages) and that the receiver does not regret his credulity. In a credible message profile the types mentioned in the messages of the profile are, by definition, telling the truth and therefore different messages mention disjoint sets of types. Types, not mentioned in any message of the profile, can do whatever they like, because their choice does not influence the receiver's choice of action. We do not assume that the messages in a profile are disjoint or that types tell the truth. This will be a consequence of the definition of a (completely) acceptable message profile as a Nash equilibrium.

(2') A second difference with Rabin's theory lies in the fact that we do not need a *message profile theory* that selects (by iterated strong dominance) the rationalizable strategy tuples from the strategy tuples 'admitted by the credible message profile'. Nash behavior is always rationalizable. We, however, select from the completely acceptable message profiles the one that gives the receiver the maximal payoff.

As 'the proof of the pudding is in eating it', we will frequently compare the predictions made by Rabin's theory with the outcomes of our method. In fact, a bit prematurely, the differences can be illustrated by two very simple examples.

**Example 1** (Rabin (1990), Example 8)

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 10 | 0 | 8 |
| $t_2$ | 6 | 8 | 0 |

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 10 | 4 | 0 |
| $t_2$ | 0 | 0 | 4 |

$p_{t_1} = p_{t_2} = \frac{1}{2}$.

In this example there are two types $t_1$ and $t_2$. The first matrix gives the payoffs to the sender, the second matrix the payoffs to the receiver. Rabin's theory predicts that type $t_1$ will always induce action $a_1$ and that type $t_2$ does also induce action $a_1$ but might try to do better with an other message. The question is, what other message? According to the theory we will develop in the next sections there are two possible message profiles, namely $\{[t_1],[t_2]\}$ and $\{[T]\}$. Under the first message profile type $t_1$ triggers (i.e. send $[t_1]$, the message is believed and the best response is chosen by the receiver) the action $a_1$ and $t_2$ triggers the action $a_3$. Both types prefer the outcome corresponding with $a_1$ and $\{[t_1],[t_2]\}$ is not a CAMP. Only the message profile $\{[T]\}$ remains. Anyhow, type $t_2$ cannot induce the receiver to use the weakly dominated action $a_2$.

**Example 2** (Rabin (1990), Example 9)

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 2 | -1 | 0 |
| $t_2$ | -1 | -2 | 0 |

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 3 | 0 | 2 |
| $t_2$ | 0 | 3 | 2 |

$p_{t_1} = p_{t_2} = \frac{1}{2}$.

Rabin claims that all types induce action $a_3$. Further he deems the message $[t_1]$ not to be a credible message. By the way, it should be noted that Rabin returns here to a (restricted) Nash behavior. We find in this example one CAMP: $\{[T]\}$ with payoff vector $[(0,0);2]$. The reader may wonder why the message $[t_1]$ is not acceptable. Sending the message $[t_1]$ will trigger the action $a_1$, if the receiver believes that the chance that type $t_1$ is sending this message is higher than $\frac{2}{3}$. So he has to find out what message type $t_2$ would have sent. As well by sending $[t_2]$ as by sending $[T]$ type $t_2$ would betray his type (if type $t_1$ sends message $[t_1]$ with high probability). So sending the message $[t_1]$ and thereby 'destroying' the credibility of this message is a good option for type $t_2$. Type $t_1$ is not able to prevent this.

# 1 The model

A sender-receiver game is a 2-person strategic game with incomplete information. Player 1 is the sender and has one of finitely many types $t \in T$. Player 1 knows his type. Player 2 is the receiver; he does not know the type of the sender but he has an a priori probability distribution $p = p_T > 0$ on the set of types $T$ which is also known by the sender. The receiver chooses an action $a$ from a finite set of actions $A$. The payoff to player 1 is $U_{t,a}$, if his type is $t$ and the action $a$ is chosen. For player 2 the payoff is $V_{t,a}$.

It may be profitable for both players, if the sender reveals some information about his type and he has a finite set $M$ of messages at his disposal to do so. Sometimes it is assumed that messages are costly, i.e. there is a cost function $c: M \to \mathbb{R}_+$. In this paper we will assume that messages are costless ($c = 0$). We collect the payoffs in two $T \times A$-matrices $U = (U_{t,a})$ and $V = (V_{t,a})$. So, the problem is given by

$$\langle p \in \Delta(T), U, V: T \times A \to \mathbb{R}, M \rangle.$$

If we model this situation, naively, as a Bayesian game, the strategy space of player 1 consists of the set of stochastic $T \times M$-matrices $X$ and the strategies of player 2 are the stochastic $M \times A$-matrices $Y$, i.e.

$$X = (X_{t,m})_{t \in T, m \in M} \text{ with } X \geq 0 \text{ and } \sum_{m \in M} X_{t,m} = 1 \text{ for all types } t$$

$$Y = (Y_{m,a})_{m \in M, a \in A} \text{ with } Y \geq 0 \text{ and } \sum_{a \in A} Y_{m,a} = 1 \text{ for all messages } m.$$

The number $X_{t,m}$ denotes the probability that type $t$ sends message $m$. The interpretation of the numbers $Y_{m,a}$ is similar.

If player 1 and player 2 play the strategies $X$ and $Y$, respectively, the stochastic $T \times A$-matrix $Z(X,Y) := X \cdot Y$ gives the probabilities that type $t$ is met by action $a$. So, type $t$ of player 1 maximizes

$$U(X, Y \mid t) := \sum_{a \in A} Z_{t,a} U_{t,a}$$

and player 2 maximizes

$$V(X,Y) := \sum_{t \in T} p(t) [\sum_{a \in A} Z_{t,a} V_{t,a}].$$

The set of Bayesian equilibria consists of the Nash equilibria of the $|T| + 1$-person 'agent normal form game'. This nonempty set inherits from his 2-person origin the property to

be the irredundant union of finitely many maximal convex (and exchangeable) subsets, the so-called *Nash components* (see Borm et al. (1996)).

For use later on, we remind the reader that the pure best responses to a strategy $X$ can be found by Bayesian inference: for a message $m \in M$ we define the conditional probability vector $\bar{X}_m$ under the condition that $X$ is played and message $m$ is received:

$$\bar{X}_m(t) := \frac{p(t)X_{t,m}}{\sum_{t\in T} p(t)X_{t,m}}.$$

The best reaction to message $m$ is an action that maximizes $a \to V(\bar{X}_m, a)$.

However, as many authors have argued (cf. a.o. Rabin (1990), Farrell (1993)), Nash equilibria do not explain how the communication between the agents takes place. To make this point clear, let us consider the following situation:

$$T = \{t_1, t_2\}, \quad A := \{a_1, a_2\}, \quad U = V = \begin{array}{c|cc} & a_1 & a_2 \\ t_1 & 1 & 0 \\ t_2 & 0 & 1 \end{array}$$

In this situation of *common interests* it is clear that type $t_1$ will try to convince player 2 to play $a_1$ and that type $t_2$ will do the same with action $a_2$. Moreover, player 2 will be easily convinced. So, *different* types must send *different* messages to communicate their types. And, in fact, the strategy $X$ in which different types send different messages together with the strategy $Y$ wherein player 2 'understands the message' and chooses the appropriate actions is a Nash equilibrium. It is however unclear how player 2 will be able to 'understand the message'. Let us assume that there are three messages called 'blue', 'red' and 'yellow' and that type $t_1$ sends message 'blue' and type $t_2$ sends message 'red'. Then player 2 has to infer that 'blue' means $t_1$ and 'red' means $t_2$ but he will not be able to do so, as it could also be the other way around that 'red' means type $t_1$ and 'blue' type $t_2$ (also forming a Nash equilibrium with the right guess of player 2). Even if we assume that sending messages is costly, these messages can not discriminate between these two symmetric types. There must be an *a priori relation between types and messages.* The most obvious solution for this problem is to see messages as 'bearers of information', they contain *information*, preferably about the types of player 1. Maybe that is the reason why in daily life people do not just send an envelope of a particular color with the appropriate number of stamps on it (a costly message) but put a letter in it:

*The medium, even if it is an expensive medium, is not the message.*

In the literature many lines are devoted to necessary conditions that make communication between agents possible. Players should 'share a meaningful sufficiently rich vocabulary'

and 'have a common understanding to interpret statements according to their literal meaning' (Farrell (1993)). We do not discuss these issues. Certainly, they are interesting, just as the observation of Rabin (1990) that 'messages should come from a common language pre-dating the specific strategic situation (...) with which the agents can richly describe all relevant strategic issues' and 'that agents, most likely, will use more natural language, such as "You should invest in my company.".' This is, however, not the subject of this paper.

In fact, we will make a short-cut by saying that the agents have a *communication channel with sufficient possibilities.* So, if player 1 uses a certain language (or smoke signals), player 2 understands the language (or the smoke signals). If player 1 uses metaphors or a code, player 2 understands the metaphors or has a decoding mechanism.

> *If player 2 does not get all the information he wants to have, the reason is not the insufficiency of the communication channel but player 1's unwillingness to give him that information.*

In the sequel we will even assume that the message space consists solely of the *unambiguous messages* $[S]$ saying "My type is one of the types in $S$" for $S \in 2^T \backslash \{\emptyset\}$. The message $[T]$ is used to convey no new information. Therefore, it is not necessary for player 1 'to babble', 'to speak gibberish' or 'to remain silent'; he can simply use the message $[T]$, in case he does not want to convey information.

After we have removed the possible insufficiency or ambiguity of the communication channel (and not earlier), the real issue of the paper emerges: the *credibility of messages.* After player 2 has found out what player 1 is telling him about his (player 1's) type, he has to decide if he can *trust* the information he has obtained. Before we come to this issue, we will answer two related questions, namely what will player 2 do, if he does not get any 'credible' information and what, if he gets the information $[S]$ and believes the message?

We assume that both players are *expected utility maximizers* and that both players are, moreover, *Bayesian players.* Accordingly, if player 2 gets no (new) credible information, he will play an action $a_T$ that maximizes $\sum_{t \in T} p(t) V_{t,a}$; if he gets the information $[S]$ and he has *no reason for doubt*, he will maximize $\sum_{t \in T} p_S(t) V_{t,a}$. By $p_S$ we mean the 'Bayesian update' of $p$ under the condition that $t \in S$ is true (believed). To make things easier we assume that, for all $S \in 2^T \backslash \{\emptyset\}$, the function $a \to \sum_{t \in T} p_S(t) V_{t,a}$ has exactly one optimal action $a_S$ and also that for each type $t \in T$ the values of $U_{t,a}$ are different. We call such a triple *generic.* Almost all triples $(p, U, V)$ satisfy these conditions, i.e. any triple $(p, U, V)$ can be made generic by an arbitrary small perturbation of $(p, U, V)$.

We call the action $a_T$ (in fact, 'always choosing $a_T$ whatever the message may be') the

*default strategy* of player 2. For player 1 the *default strategy* is 'sending the message $[T]$ whatever his type may be'. Note that the pair of default strategies form a Nash equilibrium (as deviating behavior of one of the players is not rewarded by the other player), and that the players have a common interest *to do better than the default strategy.* That is the only reason why player 1 takes the effort (and maybe the cost) to send a message and player 2 tests the messages on credibility. By transferring the information $[S]$, player 1 tries to convince player 2 to deviate from the default action $a_T$ to $a_S$. So, player 2 must formulate some *behavior rules* saying how he investigates the credibility of messages $[S]$ given the information $(p, U, V)$. These rules should be formulated *without any reference to a specific game* and should be *applicable to every specific game.* Furthermore, player 1 must know how player 2 comes to his judgement. He must know the *behavior protocol.*

So, we come to the following chronology:

$t=0$ The players get acquainted with each other's way of expressing themselves, learn a *common language* rich enough to communicate messages like "My type is a type in $S$".

$t=1$ Player 2 tells player 1 how he will form his opinion about the credibility of messages, his behavior protocol. Here the common language is challenged more seriously. One may fear that smoke signals are no longer adequate.

$t=2$ Player 1 and 2 learn the type space $T$, the action space $A$, the payoff matrices $U$ and $V$ and the a priori probability distribution $p$. Here the common meeting ends.

$t=3$ Player 1 is informed *privately* about his type. Player 2 knows that this happens.

$t=4$ Player 1 sends a message $[S]$ from $M = 2^T \backslash \{\emptyset\}$ and player 2 receives the message and learns that player 1 tries to convince him to deviate from the default action $a_T$ to $a_S$.

$t=5$ Player 2 forms his opinion about the credibility of $[S]$ by using the earlier communicated decision rules (see $t = 1$).

$t=6$ and acts accordingly.

Stage $t = 0$ may be deleted, if the players know that they will understand each other, and stage $t = 6$ is an automatism. We will use the word 'trigger' or 'induce' for the combination of the steps $t = 5$ and $t = 6$. The remaining issue is the formulation of player 2's behavior protocol that he tells player 1 in stage $t = 1$, that player 1 will use to select his message in

stage $t = 4$ and that player 2 will use in stage $t = 5$.

A part of the protocol has been written already, namely that both players are *Bayesian players who maximize expected utility.* Being a paradigm in game theory and in many economic theories, we assume that the players do not have any problem to believe these statements. Notice, however, that also the main part of the protocol, how to form an opinion about the credibility of messages, should be *convincing for both players.* In particular, player 1 should be convinced that player 2 will really apply his protocol in stage $t = 5$. Then he can use it to anticipate player 2's behavior in stage $t = 4$. To make anticipation possible it is necessary that player 1 knows that player 2 accepts or does not accept a message as a credible message and not something in between. The behavior protocol should not introduce a new kind of ambiguity.

**Remark 1** Up to now we called accepted messages *credible.* But, in fact, player 2 is not interested in the *truth of a message* (that a type sending a message $[S]$ is indeed a member of $S$) but that it is *profitable* for him to accept $[S]$ and to switch from the default action $a_T$ to action $a_S$. He is an *expected utility maximizer* and not a *searcher for truth.* Therefore, we will call messages that pass player 2's tests *acceptable* instead of *credible.*

# 2 Acceptable message profiles

In Rabin (1990) a formulation of a possible behavior protocol can be found. It is called a *credible message profile.* The author puts, however, severe restrictions on credible messages with the consequence that in many examples where you would expect some meaningful communication, this turns out to be impossible. The main property of a credible message (profile) that the author requires, is the optimality of the chosen action $a_S$ for all types mentioned in the message $[S]$. In our opinion this is a too severe restriction. If both players can *gain with respect to the default equilibrium payoff*, they may also have reasons to send and to believe some messages. This is the idea we will try to elaborate in the present paper. With the previously mentioned papers it will have in common that the acceptability of messages can not be derived from the message alone but only from the message as a member of a family of other acceptable messages. The reason to call a message acceptable is partially found in the presence of other acceptable messages. To avoid even the slightest suspicion of an infinite regress we introduce the concept of an *acceptable message profile.*

Before we come to the definition of an acceptable message profile, we introduce the following notation:

$a \succeq_t \bar{a}$ if $U_{t,a} \geq U_{t,\bar{a}}$ and $a \succeq_t^* \bar{a}$ if $V_{t,a} \geq V_{t,\bar{a}}$.

The meaning is: if $a \succeq_t \bar{a}$ and player 1 has type $t$, he weakly prefers action $a$ to action $\bar{a}$; if $a \succeq_t^* \bar{a}$ and player 2 thinks that the type of player 1 is $t$, he weakly prefers action $a$ to action $\bar{a}$. As we assume that $(p, U, V)$ are generic, the relations $\succeq_t$ and $\succeq_t^*$ are asymmetric.

Let $\mathcal{C}$ be any non-empty collection of messages in $M$. Then we define a strategy for player 2, $\mathrm{Acc}_{\mathcal{C}} : M \to A$ (accept the messages in $\mathcal{C}$ and no other), by

$$\mathrm{Acc}_{\mathcal{C}}([S]) := \begin{cases} a_S & \text{if } [S] \in \mathcal{C} \\ a_T & \text{if } [S] \notin \mathcal{C} \end{cases} .$$

If player 1 knows that player 2 will accept the messages from $\mathcal{C}$ and none of the messages outside $\mathcal{C}$, each of his types will send a message $[S]$ that triggers the best action ($a_S$) of the actions that can be induced by a message from $\mathcal{C}$ or by a *not acceptable message*. We call such a strategy $\mathrm{Rev}_{\mathcal{C}} : T \to M$ (reveal your type according to $\mathcal{C}$):

$$\mathrm{Rev}_{\mathcal{C}}(t) = [S] \quad \Longrightarrow \quad \begin{cases} a_S \succeq_t a_{\bar{S}} \text{ for all } [\bar{S}] \in \mathcal{C} \text{ and } a_S \succeq_t a_T & \text{if } [S] \in \mathcal{C} \\ a_T \succeq_t a_{\bar{S}} \text{ for all } [\bar{S}] \in \mathcal{C} & \text{if } [S] \notin \mathcal{C} \end{cases} .$$

Then each of the strategies $\mathrm{Rev}_{\mathcal{C}}$ is, by definition, a best response to $\mathrm{Acc}_{\mathcal{C}}$.

The following example shows that $\mathrm{Rev}_{\mathcal{C}}$ may consist of more than one strategy.

**Example 3**

| | $a_0$ | $a_1$ |
|---|---|---|
| $t_1$ | 0 | 1 |
| $t_2$ | 0 | 1 |
| $t_3$ | 0 | 1 |

| | $a_0$ | $a_1$ |
|---|---|---|
| $t_1$ | 3 | 1 |
| $t_2$ | 3 | 6 |
| $t_3$ | 3 | 1 |

$p_{t_i} = \frac{1}{3},\ i = 1, 2, 3.$

All types of player 1 prefer action $a_1$ to the default action $a_0$. The messages $[t_1, t_2]$, $[t_2]$ and $[t_2, t_3]$ are the messages that trigger action $a_1$. If these messages are elements of a message profile $\mathcal{C}$, every strategy in which each type sends one of these messages is an element of $\mathrm{Rev}_{\mathcal{C}}$.

In fact, we restrict the domain of $\mathrm{Rev}_{\mathcal{C}}$ to strategies in which types who want the *same* action, send the *same* message, i.e. if $\mathrm{Rev}_{\mathcal{C}}(t) = [S]$ and $\mathrm{Rev}_{\mathcal{C}}(t') = [S']$ with $a_S = a_{S'}$, then $[S] = [S']$. So, in Example 3 all types send message $[t_1, t_2]$ or $[t_2]$ or $[t_2, t_3]$. The idea is that each message triggering action $a$ is only sent with the purpose to induce $a$. There is no additional information sent. Or, to say it differently each message $[S]$ will be considered to contain the information "play $a_S$" and not to contain any *residual* information. In the example the message $[t_2]$ can be sent by each of the types. *Truth telling is not focal, if this implies residual information* e.g. my type is with high probability $t_2$, if $[t_2]$ is sent in Example 3.

**Definition 2.1** A collection $\mathcal{C}$ is called an *acceptable message profile*, if $\mathrm{Acc}_{\mathcal{C}}$ is also a best response to the strategies $\mathrm{Rev}_{\mathcal{C}}$, i.e. if $(\mathrm{Rev}_{\mathcal{C}}, \mathrm{Acc}_{\mathcal{C}})$ is a Nash equilibrium.

Although this concept does not solve the communication problem between the players, it is at least a first stepping stone in the uncertain world of communication. If player 2 believes the messages from $\mathcal{C}$ and no other messages and player 1 plays a best response to that strategy, player 2 will, after all, be justified to have believed the messages from $\mathcal{C}$. Acceptable message profiles are viable ways of communication, once the players can agree upon the choice of an acceptable message profile without any further communication. We will address this important issue later on. First we will investigate the merits of an acceptable message profile.

The following proposition shows the remarkable fact that every acceptable message profile generates an acceptable message profile $T(\mathcal{C})$ with the same payoffs as $\mathcal{C}$ in which every type is 'telling the truth' and the messages mention disjoint sets of types.

If $\mathcal{C}$ is a message profile and $a$ is an action from $A$ we define $T(a\,|\,\mathcal{C})$ to be the set of types $\{t : \mathrm{Rev}_{\mathcal{C}}(t)$ triggers the action $a\}$. Let $A_{\mathcal{C}}$ be the set of actions $a$ with $T(a\,|\,\mathcal{C}) \neq \emptyset$. The message profile $T(\mathcal{C})$ consists of the messages $[T(a\,|\,\mathcal{C})]$ for all $a \in A_{\mathcal{C}}$. Note that $T(a\,|\,\mathcal{C})$ is not dependent on the choice of $\mathrm{Rev}_{\mathcal{C}}$.

**Proposition 2.2** *(a) Let $\mathcal{C}$ be any message profile. Then the best responses to any $\mathrm{Rev}_{\mathcal{C}}$ are the strategies with $[S] \to a_{T(a_S\,|\,\mathcal{C})}$ if $a_S \in A_{\mathcal{C}}$.*
*(b) $\mathcal{C}$ is an acceptable message profile if and only if $a_S = a_{T(a_S\,|\,\mathcal{C})}$ for all messages $[S] \in \mathcal{C}$ with $a_S \in A_{\mathcal{C}}$.*
*(c) If $\mathcal{C}$ is an acceptable message profile, $T(\mathcal{C})$ is also an acceptable message profile with the same payoffs for both players. The acceptable message profile $T(\mathcal{C})$ has the additional properties that $t$ sends message $[T(a\,|\,\mathcal{C})]$ if and only if $t \in T(a\,|\,\mathcal{C})$ ('truth telling') and that different messages mention disjoint sets of types.*

**Proof** (a) If $\mathrm{Rev}_{\mathcal{C}}$ is played by the sender and player 2 receives the message $[S]$, his updated belief over the types is $p_{T(a_S\,|\,\mathcal{C})}$ and, by definition, $a_{T(a_S\,|\,\mathcal{C})}$ is the best reaction to that belief. The reaction to messages never sent is immaterial.
(b) This follows immediately from (a).
(c) The types sending the message $[S]$ in $\mathrm{Rev}_{\mathcal{C}}$, will send the message $[T(a_S\,|\,\mathcal{C})]$ in $\mathrm{Rev}_{T(\mathcal{C})}$, as $a_S$ is also induced by the message $[T(a_S\,|\,\mathcal{C})]$ and the actions that can be triggered by any message from $T(\mathcal{C})$ is a subset of $\{a_S : [S] \in \mathcal{C}\} \cup \{a_T\}$. The reaction of the receiver to $[S]$ in $\mathrm{Acc}_{\mathcal{C}}$, namely $a_S$, is the same as the reaction of the receiver to $[T(a\,|\,\mathcal{C})]$ in $\mathrm{Acc}_{T(\mathcal{C})}$, because $(\mathrm{Rev}_{\mathcal{C}}, \mathrm{Acc}_{\mathcal{C}})$ is an equilibrium (apply (b)). Then it is clear

from (a) that $\mathrm{Acc}_{T(\mathcal{C})}$ is a best response to $\mathrm{Rev}_{T(\mathcal{C})}$. $\triangleleft$

From Proposition 2.2, (c) follows that we can restrict the attention to acceptable message profiles in which all types tell the truth and different messages mention disjoint sets of types. Every payoff vector generated by an acceptable message profile can also be generated by an acceptable message profile with these additional properties.

**Definition 2.3** We call acceptable message profiles in which all types tell the truth and different messages never contain the same type *completely acceptable message profiles* (CAMP).

Here we show a slight preference for truth telling:

> *If the same result can be reached by being honest as well as by being dishonest, we assume that players prefer honesty.*

An additional advantage is that there are less CAMP's and it may be easier to find all of them.

In the next proposition we prove that in antagonistic games (i.e. if $a \succeq_t b$ is equivalent to $b \succeq_t^* a$ for all $t \in T$ and all $a, b \in A$), only the message profile $\mathcal{C} := \{[T]\}$ is completely acceptable. For common interest games (i.e. if $a \succeq_t b$ is equivalent to $a \succeq_t^* b$ for all $t \in T$ and all $a, b \in A$), there is only one completely acceptable message profile in which all types reveal all relevant information. This is not necessarily the 'separating' message profile $\mathcal{C}_0 := \{[t] : t \in T\}$, since different types may have the same most preferred action.

**Proposition 2.4** *(a) In an antagonistic game only the message profile* $\{[T]\}$ *is completely acceptable.*
*(b) In a common interest game there is exactly one completely acceptable message profile that guarantees all types of player 1 their highest payoff.*

**Proof** (a) Let $(U, V)$ be the payoff matrices in a generic antagonistic game. We prove that no profile $\mathcal{C}$ with more than one element is completely acceptable. Let $[S]$ and $[S']$ be two different (and therefore disjoint) messages from a completely acceptable message profile $\mathcal{C}$. If $a_S = a_{S'}$, then $S \cup S' \subseteq T(a_S \,|\, \mathcal{C})$ and all types in $S \cup S'$ send the same message. Then the types in $S$ or the types in $S'$ do not tell the truth. If $a_S \neq a_{S'}$, the types in $S$ prefer $a_S$ to $a_{S'}$ and for the types in $S'$ the opposite preference holds. For player 2 the preferences are opposite and therefore, it is better for player 2 to respond $[S]$ with $a_{S'}$ and $[S']$ with $a_S$. The strategy $\mathrm{Acc}_{\mathcal{C}}$ is not a best response.
(b) Every type $t$ has exactly one best action $a(t)$. Collect the types with the same best

action as $t$ in a set $R(t)$. Consider the message profile $\mathcal{C} := \{[R(t)] : t \in T\}$. Then the messages in $\mathcal{C}$ induce a partition on $T$, the message $[R(t)]$ triggers the action $a(t)$, i.e. $a_{R(t)} = a(t)$ (by common interest) and therefore the types in $R(t)$ send the message $[R(t)]$ to obtain their highest payoff. Therefore, $T(\mathcal{C}) = \mathcal{C}$. So, the message profile $\mathcal{C}$ is completely acceptable.

Let $\mathcal{C}'$ be a different completely acceptable message profile also triggering the best action for each type and $[R]$ is a message from $\mathcal{C}'$. All types in $R$ must have the same best action and therefore $R \subseteq R(t)$ for some type $t$. If there is no equality, there is a type in $R(t)$, triggering the action $a(t)$ by means of a different message $[R'] \in \mathcal{C}'$. Then also $R' \subseteq R(t)$. So, $R \cup R' \subseteq T(a(t) \,|\, \mathcal{C}')$ and the types in $R \cup R'$ send the same message in $\text{Rev}_{\mathcal{C}'}$. Then some type is not telling the truth. ◁

# 3 Some examples and the behavior protocol.

In the following examples we compute completely acceptable message profiles in a rather ad hoc way. In Section 4 we will do it in a more systematic way.

We assume, in all examples, that all types have the same probabilities. We only give the $U$- and $V$-matrix. If a payoff has an upper index $*$, it means that the payoff is slightly larger than the given number (to make the example generic). Most of the examples are from Rabin (1990).

**Example 4** (Rabin (1990), Example 1)

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 10 | $0^*$ | 0 |
| $t_2$ | 0 | 10 | 5 |
| $t_3$ | 0 | 10 | 5 |

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | $10^*$ | 0 | 0 |
| $t_2$ | 0 | 10 | 7 |
| $t_3$ | 0 | 0 | 7 |

For the moment we assume that all messages will be believed.

$$t_1 : a_1 \succ a_2 \succ a_3 \qquad [t_1], [t_1, t_2], [t_1, t_3] \longrightarrow a_1$$
$$t_2 : a_2 \succ a_3 \succ a_1 \qquad [t_2] \longrightarrow a_2$$
$$t_3 : a_2 \succ a_3 \succ a_1 \qquad [t_3], [t_2, t_3], [T] \longrightarrow a_3$$

The first block gives the preferences of the types, the second block the actions 'triggered' by the messages. The first inconsistency with a completely acceptable message profile lies in the fact that $t_3$ will send message $[t_2]$ to trigger his most preferred action $a_2$. He will do so, as long as $[t_2]$ is available. So, we have to delete $[t_2]$. So, under truth-telling type $t_2$ can only send $[t_2, t_3]$ or $[T]$ to trigger his second best action $a_3$ and type $t_3$ can do the same.

Then type $t_1$ will send $[t_1]$ and the types $t_2$ and $t_3$ will send truthfully the message $[t_2, t_3]$. Both of them could also send $[t_3]$ or $[T]$. This gives an acceptable but not a completely acceptable message profile. There are two completely acceptable message profiles, namely $\mathcal{C}_0 := \{[T]\}$ and $\mathcal{C}_1 := \{[t_1], [t_2, t_3]\}$. The latter one is more profitable for both players. Rabin comes to the conclusion that the message $[t_1]$ is credible and that the message $[t_2]$ is incredible. The credibility of $[t_2, t_3]$ is not commented upon. We think Rabin would reject this message because some type does not get his highest payoff.

**Example 5** (Rabin (1990), Example 2)

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 1 | -2 | 0 |
| $t_2$ | -2 | -1 | 0 |

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $t_1$ | 3 | 0 | 2 |
| $t_2$ | 0 | 3 | 2 |

$$t_1: \quad a_1(\leftarrow [t_1]) \succ a_3(\leftarrow [T]) \succ a_2(\leftarrow [t_2])$$
$$t_2: \quad a_3(\leftarrow [T]) \succ a_2(\leftarrow [t_2]) \succ a_1(\leftarrow [t_1])$$

If message $[T]$ is in the message profile, type $t_2$ will send this message. If we delete $[T]$, we get the message profile $\mathcal{C}_1 := \{[t_1], [t_2]\}$. This is not a completely acceptable message profile, since type $t_2$ will send the not-acceptable message $[T]$. Rabin's theory predicts that the types will reveal themselves, even when some types would prefer that no such revelation takes place. We come to the opposite conclusion: type $t_2$ would not allow type $t_1$ to reveal himself or at least player 2 has reasons to mistrust message $[t_1]$ too.

**Example 6** (Rabin (1990), Example 5)

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|---|
| $t_1$ | -1 | 7 | 6 | 0 |
| $t_2$ | -1 | 6 | 7 | 0 |
| $t_3$ | -1 | 0* | 0 | 6 |

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|---|
| $t_1$ | 5 | 6* | 7 | 0 |
| $t_2$ | 5 | 7 | 6 | 0 |
| $t_3$ | 5 | 0* | 0 | 6 |

We give the preferences of the types and the messages that trigger these actions:

$$t_1: \quad a_1(\leftarrow [t_2], [t_1, t_2]) \succ a_2(\leftarrow [t_1]) \succ a_3(\leftarrow [t_3]) \succ a_0(\leftarrow [t_1, t_3], [t_2, t_3], [T])$$
$$t_2: \quad a_2(\leftarrow [t_1]) \succ a_1(\leftarrow [t_2], [t_1, t_2]) \succ a_3(\leftarrow [t_3]) \succ a_0(\leftarrow [t_1, t_3], [t_2, t_3], [T])$$
$$t_3: \quad a_3(\leftarrow [t_3]) \succ a_1(\leftarrow [t_2], [t_1, t_2]) \succ a_2(\leftarrow [t_1]) \succ a_0(\leftarrow [t_1, t_3], [t_2, t_3], [T])$$

The messages $[t_1]$ and $[t_2]$ will used by the types $t_2$ and $t_1$, respectively. Both messages must be deleted to obtain a completely acceptable message profile. So, the best the sender can get is action $a_1$ for the types $t_1$ and $t_2$ by sending the message $[t_1, t_2]$ and action $a_3$ for type $t_3$ by message $[t_3]$. Rabin concludes that the types strongly prefer to reveal if the type is $t_3$ or not. This is exactly what our theory predicts.

**Example 7** (Rabin (1990), Example 10)

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ |
|---|---|---|---|---|---|---|---|---|---|
| $t_1$ | 0 | 10 | 0* | 0 | 5 | -1 | -1 | -1 | -1 |
| $t_2$ | 0 | 5 | 10 | 0* | 0 | -1 | -1 | -1 | -1 |
| $t_3$ | 0 | 0 | 5 | 10 | 0* | -1 | -1 | -1 | -1 |
| $t_4$ | 0 | 0* | 0 | 5 | 10 | -1 | -1 | -1 | -1 |

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ |
|---|---|---|---|---|---|---|---|---|---|
| $t_1$ | 3 | 4 | 0 | 0 | 4 | 5 | 0 | 0 | 0 |
| $t_2$ | 3 | 4 | 4 | 0 | 0 | 0 | 5 | 0 | 0 |
| $t_3$ | 3 | 0 | 4 | 4 | 0 | 0 | 0 | 5 | 0 |
| $t_4$ | 3 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 5 |

$$
\begin{aligned}
t_1: &\quad a_1(\leftarrow [t_1,t_2]) \succ a_4(\leftarrow [t_1,t_4]) \succ \dots \\
t_2: &\quad a_2(\leftarrow [t_2,t_3]) \succ a_1(\leftarrow [t_1,t_2]) \succ \dots \\
t_3: &\quad a_3(\leftarrow [t_3,t_4]) \succ a_2(\leftarrow [t_2,t_3]) \succ \dots \\
t_4: &\quad a_4(\leftarrow [t_1,t_4]) \succ a_3(\leftarrow [t_3,t_4]) \succ \dots
\end{aligned}
$$

If we want to keep the message $[t_1, t_2]$ in a completely acceptable message profile, the messages $[t_2, t_3]$ and $[t_1, t_4]$ must be removed and we find the completely acceptable message profile $\mathcal{C}_1 := \{[t_1, t_2], [t_3, t_4]\}$. Also the message profile $\mathcal{C}_2 := \{[t_2, t_3], [t_1, t_4]\}$ is completely acceptable. The union of these two profiles, however, is not acceptable. If $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ is used for communication, $\mathrm{Acc}_{\mathcal{C}}$ is not a best response to $\mathrm{Rev}_{\mathcal{C}}$: $T(\mathcal{C}) = \{[t_1], [t_2], [t_3], [t_4]\}$. Here we see a quite serious problem to deal with. There is more than one CAMP and some types, namely $t_1$ and $t_3$, prefer to use message profile $\mathcal{C}_1$ and the other types, $t_2$ and $t_4$, prefer message profile $\mathcal{C}_2$. Rabin believes that the fully pooling equilibrium is a very plausible equilibrium in this game. Our conclusion is different, for the moment. Although there are problems (see Remark 2), the players should at least try to coordinate on one of the two CAMP's.

All these examples give us an indication for a senseful behavior rule for player 2.

> *No message $[S]$ will be accepted unless there is a completely acceptable message profile containing $[S]$.*

**Remark 2** If a message $[S]$ is a member of two different completely acceptable message profiles, the reaction of player 2 will be the same. So, it is not important for player 2 to know which completely acceptable message profile player 1 has in mind but it is important for him to know that all types of player 1 use the same completely acceptable message profile. He has, however, no means to check this and different types have an incentive to use different (completely acceptable) message profiles. In Example 7 e.g. we have seen that the types $t_1$ and $t_3$ prefer $\mathcal{C}_1$ and that the types $t_2$ and $t_4$ like to use $\mathcal{C}_2$.

Because the receiver determines the behavior protocol, he is not willing to coordinate on a certain message profile if there is another one that gives him a better payoff. He is only willing to coordinate on a CAMP if there is no other CAMP which is better (in payoff) for him. We will call such a CAMP *maximal*.

Note that most triples $(p, U, V)$ admit only one maximal CAMP, meaning that every triple with more than one maximal CAMP can be transformed into a triple with only one maximal CAMP by an arbitrarily small perturbation of $V$. So, having only one maximal CAMP is a generic property.

Summarizing the assumptions made with respect to the behavior of both players we come to the following *behavior protocol*:

*(1) Both players maximize expected utility and are Bayesian players.*

*(2) Player 2 will accept a message if and only if it is a member of the maximal completely acceptable message profile.*

*(3) If a message $[S]$ is accepted, player 2 will choose action $a_S$; if a message is not accepted, the default action $a_T$ is chosen. So, messages triggering the same action are considered to be equivalent; there is no residual information.*

In order to handle the behavior rules adequately both players must be able to determine all (maximal) completely acceptable message profiles. How they can find them is the subject of the next section.

# 4 Determining completely acceptable message profiles.

In this section we will show how the players can determine the completely acceptable message profiles. We start with an undirected graph. The node set $N$ consists of messages $[S]$ with $a_S \succeq_t a_T$ for all $t \in S$. If a message $[S]$ contains a type $t$ with $a_T \succ_t a_S$, such a type prefers to send any not acceptable message (triggering $a_T$) to $[S]$. We connect two different messages $[S]$ and $[S']$ if the messages are 'compatible' in the sense that they can occur together in the same CAMP. There are two conditions to be satisfied. The first condition is $S \cap S' = \emptyset$. This implies e.g. that the message $[T]$ cannot be connected with any other message. The second condition that connected messages must satisfy is that none of them undermines the credibility of the other, where we say that *message* $[S']$ *undermines (the credibility of) message* $[S]$ if there is a type $t \in S$ such that $a_{S'} \succeq_t a_S$. In

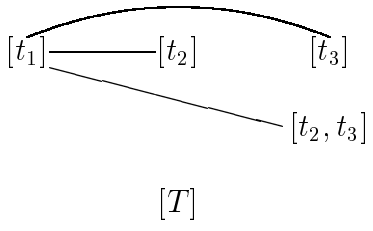
the graph obtained in this way we look for *maximal cliques*. A clique is a subset of nodes of which *each pair* is connected. They form a complete subgraph. A clique is *maximal* if it cannot be extended to a larger clique.

**Proposition 4.1** *If the messages in a clique form a partition of $T$, they form a completely acceptable message profile and conversely every completely acceptable message profile is a maximal clique in the graph just defined.*

**Proof** In a completely acceptable message profile the messages are disjoint, there is no type $t$ occurring in a message $[S] \in \mathcal{C}$ with $a_T \succ_t a_S$ or a pair of message $[S]$ and $[S']$ such that $a_{S'} \succ_t a_S$ for some type $t \in S$. We must also exclude $a_{S'} \sim_t a_S$ for a type $t \in S$. As $(p, U, V)$ are supposed to be generic, we find $a_{S'} = a_S$. Under $\mathrm{Rev}_{\mathcal{C}}$ types triggering the same action are supposed to send the same message. Then there are types in $S$ or in $S'$ not telling the truth.

Conversely, we must prove that a maximal clique that forms a partition gives a completely acceptable message profile. Suppose the messages $[S_1]$, $[S_2]$, $\ldots$, $[S_q]$ form a maximal clique and $\bigcup_{i=1}^{q} S_i = T$. For all types $t$ in $S_i$ we have $a_{S_i} \succ_t a_{S_j}$ for $j \neq i$ ($S_j$ does not undermine $S_i$) and $a_{S_i} \succeq_t a_T$ (otherwise $[S_i]$ was not a node of the graph). Let $\mathcal{C}$ be the message profile consisting of the messages $[S_i]$ for $i = 1, \ldots, q$. Note that different messages trigger different actions. At most one of these actions is $a_T$. If none of the messages trigger $a_T$, every type $t \in S_i$ sends the message $[S_i]$ and $T(\mathcal{C}) = \mathcal{C}$. Then $\mathcal{C}$ is an acceptable message profile by Proposition 2.2 and every type tells the truth. Since $\mathcal{C}$ is also a partition of $T$, it is a completely acceptable message profile. If $a_{S_i} = a_T$ for an index $i$, then all types in $S_i$ strictly prefer the default action $a_T$ to all attainable actions $a_{S_j}$ ($j \neq i$). Under $\mathrm{Rev}_{\mathcal{C}}$ they will all send message $[S_i]$ or the same not acceptable message. Whatever they send, the receiver will response with $a_T$ ($= a_{S_i}$). By consequence, $(\mathrm{Rev}_{\mathcal{C}}, \mathrm{Acc}_{\mathcal{C}})$ is a Nash equilibrium. So, also in this case we find a completely acceptable message profile. ◁

In Example 4 the nodes of the graph are the messages $[t_1]$, $[t_2]$, $[t_3]$, $[t_2, t_3]$ and $[T]$, because $a_T = a_3 \succ_{t_2} a_1 = a_{\{t_1,t_2\}}$ and $a_T = a_3 \succ_{t_2} a_1 = a_{\{t_1,t_3\}}$. Further, $a_{\{t_2\}} = a_2 \succ_{t_3} a_3 = a_{\{t_3\}}$, which implies that in the graph the node $[t_2]$ is not connected with the node $[t_3]$. So, we come to the following graph:




$[T]$

We see that the graph contains two maximal cliques covering $T$, namely $\mathcal{C}_0 = \{[T]\}$ and $\mathcal{C}_1 = \{[t_1], [t_2, t_3]\}$. Note that these collections are the same as found earlier in an ad hoc way. When $\mathcal{C}_0$ is used the (expected) payoff becomes $[(0, 5, 5); \frac{14}{3}]$; when $\mathcal{C}_1$ is used the (expected) payoff becomes $[(10, 5, 5); \frac{24}{3}]$. Obviously, $\mathcal{C}_1$ gives player 2 a larger payoff than $\mathcal{C}_0$. So, $\mathcal{C}_1$ is the unique maximal CAMP.

# References


Borm, P., Garcia-Jurado, I., Potters, J. and Tijs, S. (1996), "An Amalgamation of Games", *European Journal of Operations Research, 89*, 570–580.

Cho, I.-K. and Kreps, D.M. (1987), "Signaling Games and Stable Equilibria", *The Quarterly Journal of Economics, 102*, 179–221.

Crawford, V.P. and Sobel, J. (1982), "Strategic Information Transmission", *Econometrica, 50*, 1431–1452.

Damme, E.E.C. van (1989). "Stable Equilibria and Forward Induction", *Journal of Economic Theory, 48*, 476–496.

Farrell, J. (1993), "Meaning and Credibility in Cheap-Talk Games", *Games and Economic Behavior, 5*, 514–531.

Hurkens, S. (1995), *Games, Rules and Solutions*, Ponsen & Looijen B.V., Wageningen.

Rabin, M. (1990), "Communication between Rational Agents", *Journal of Economic Theory, 51*, 144–170.

Zapater, I. (1993), "Generalized Communication between Rational Agents", *mimeo*, Brown University, Providence, Rhode Island.